# Semantic Web for Earth and Environmental Terminology (SWEET)

Rob Raskin
Jet Propulsion Laboratory
4800 Oak Grove Dr
Pasadena, CA 91109

*Abstract*- **The Semantic Web for Earth and Environmental Terminology (SWEET) is a prototype for improving the discovery and use of Earth science data, through software understanding of the semantics of web resources. The semantic understanding is aided by the use of ontologies, or formal representations of technical concepts and their interrelations in a form that supports domain knowledge. The ultimate vision of the semantic web is of web pages with XML-type tags around ontology terms, enabling search tools to ascertain their meanings by examining the (third-party) ontology content. Such a scenario both reduces the number of false hits (where a search returns alternative, unintended meanings of a term) and increases the number of successful hits (where searcher and information provider have a syntax mismatch of the same concept).**

**For SWEET, we developed a collection of ontologies in the DAML+OIL ontology language that included both orthogonal concepts (space, time, Earth realms, physical quantities, etc.) and integrative science knowledge concepts (phenomena, events, etc.). We developed a search tool that finds alternative search terms (based on the semantics) and redirects the expanded set of terms to the GCMD Search Tool. We also explored "virtual tag" creation by search engines during the indexing process, based upon inference of meaning using natural language processing algorithms.**

## I. Introduction

Web searches for Earth science data and information are commonly hindered by syntax mismatches between information user and information provider. If the user does not enter the "correct" search terms, either not enough or too many hits are returned. The underlying cause is a lack of "common sense knowledge" inherent in the search tool. An emerging solution to this problem is through the "Semantic Web", an extension to the existing WWW environment, coordinated by W3C. The Semantic Web encodes common sense knowledge directly into web pages themselves, using broadly agreed upon namespaces and ontologies to define terms and their mutual relationships. Several such ontologies have been developed, but they lack specialized knowledge of Earth science terminology.

The motivation for our task is to improve semantic understanding of web resources by software tools, with specific application to discovery and use of Earth science data. Semantic understanding of text by automated tools is enabled through the combined use of *i)* ontologies and *ii)* software tools that can interpret the ontologies. An ontology is a formal representation of technical concepts and their interrelations in a form that supports domain knowledge. Generally, an ontology is hierarchical, with child concepts having explicit properties to specialize their parent concept(s).

If terms on web pages are tagged with the namespace of an ontology, a search tool potentially can use these metadata tags to distinguish different uses of the same term (e.g. “fall” as a season vs. “fall” as a downward motion) to eliminate false hits. It also can locate resources without having an exact keyword match, because terms such as “El Nino” have an equivalent definition in terms of its defining scientific components.

Our primary task was ontology development specific to the Earth and environmental sciences and supporting areas. We created a common sense knowledge base of the Earth sciences by developing an extension to an existing ontology using the DAML+OIL ontology language. We used this ontology in a prototype search tool that improves performance by creating additional relevant search terms based on the underlying semantics.

## II. Ontology Development

An ontology is a formal representation of technical concepts and their interrelations in a form that supports domain knowledge. Generally, an ontology is hierarchical, with child concepts having explicit properties to specialize their parent concept(s). Thus, “hydrosphere” is the parent concept of “surface water”, which is a parent of “river”, which is a parent of “Mississippi River”, etc. In this paper, we describe our experiences with development of Earth and environmental science ontologies. Our starting point of reference for ontology development was the collection of keywords in the NASA Global Change Master Directory (GCMD) [1]. This collection includes both controlled and uncontrolled keywords.

*Controlled keywords:* GCMD includes approximately 1000 controlled Earth science keywords, represented in a *taxonomy*. A taxonomy is a subject classification, as used by libraries or clearinghouses to assign a classification to a resource. In a taxonomy, properties are not passed on from parent to child, making this structure less suitable for knowledge representation purposes. Several hundred additional controlled keywords are defined for datasets and

data services, including instruments, data centers, missions, etc.

*Free-text terms:* Over 20,000 uncontrolled keywords have been submitted by data providers. These terms tend to be more general than or synonymous with the controlled terms. Examples of frequently submitted terms include: climatology, remote sensing, EOSDIS, statistics, marine, geology, vegetation, etc.

We used the GCMD controlled keywords as an initial guide in developing our ontologies, although we made several changes to the keyword structure. For example, rather than define a compound concept such as *air temperature*, we separated the physical property (*temperature*) from the element that the property applies to (*air*). This provides a more scalable solution to a growing knowledge base. Knowledge of the independent concepts of *"air"* and "*temperature"* provide a complete understanding of "*air temperature"* without a need to create an explicit definition of the compound concept.

From our experience with ontology development, we concluded that the following guiding principles are essential:

1. **Scalability:** An ontology should be easily extendable to enable specialized domains to build upon more general ontologies already generated.
2. **Application-independence:** The structure and contents of an ontology should be based upon the inherent knowledge of the discipline, rather than on how the domain knowledge is used.
3. **Natural language-independence:** The structure should provide a representation of *concepts*, rather than of terms. The concepts remain the same regardless of the inclusion of slang, technical jargon, foreign languages, etc. Synonymous terms (e.g., marine, ocean, sea, oceanography, ocean science) can be mapped separately to an ontology element
4. **Orthogonality:** Compound concepts should be decomposed into their component parts, to make it easy to recombine concepts in new ways.
5. **Community involvement:** Community input should guide the development of any ontology.

## III. SWEET ONTOLOGIES

SWEET includes the following collection of ontologies:

*Earth Realm*

The "spheres" of the Earth constitute an *EarthRealm* ontology, based upon the physical properties of the planet. Elements of this ontology include "atmosphere", "ocean", and "solid earth", and associated subrealms (such as "ocean floor" and "atmospheric boundary layer"). The subrealms generally are distinguished from their parent classes, based on the property of altitude, e.g., "troposphere" is the subclass of "atmosphere" where elevation is between 0 and 15 km.

*Non-Living Element*

This ontology includes the non-living building blocks on nature, such as: particles, electromagnetic radiation, and chemical compounds.

*Living Element*

This ontology includes plant and animal species. It was imported from the "biosphere" taxonomy of GCMD.

*Physical Property*

A separate ontology was developed for physical properties that might be associated with any component of *EarthRealm, NonLivingElements, or LivingElements. PhysicalProperties* include "temperature", "pressure", "height", "albedo", etc.

*Units*

Units are defined using Unidata's UDUnits. The resulting ontology includes conversion factors between various units. Prefixed units such as km are defined as a special case of m with appropriate conversion factor.

*Numerical Entity*

Numerical extents include: interval, point, 0, $\mathbf{R}^2$, …
Numerical relations include: greaterThan, max, …

*Temporal Entity*

Time is essentially a numerical scale with terminology specific to the temporal domain. We developed a time ontology in which the temporal extents and relations are special cases of numeric extents and relations, respectively.
Temporal extents include: duration, season, century, 1996, …
Temporal relations include: after, before, …

*Spatial Entity*

Space is essentially a 3-D numerical scale with terminology specific to the spatial domain. We developed a space ontology in which the spatial extents and relations are special cases of numeric extents and relations, respectively.
Spatial extents include: country, Antarctica, equator, inlet, …
Spatial relations include: above, northOf, …

*Phenomena*

A phenomena ontology is used to define transient events. A phenomenon crosses bounds of other ontology elements. Examples include: hurricane, earthquake, El Nino, volcano, terrorist event, and each has associated *Time*, *Space*, *EarthRealm*s, *NonLivingElements*, *LivingElements*, etc. We also include specific instances of phenomena, spanning approximately 50 events over the past two decades.

Human Activities

This ontology is included for representing impacts of environmental phenomena such as commerce, fisheries, etc.

## IV. Ontologies As A Unifying Knowledge Framework

Most of the above ontology categories represent orthogonal concepts. Each of these orthogonal dimensions constitutes a hierarchy of complexity (or richness); traversing down the associated tree follows the path of reductionism by adding additional details to more abstract concepts. An additional dimension "phenomena" is synergetic rather than orthogonal to the others. The phenomena entries describe synthesizing concepts that utilize elements from the other ontologies (e.g., a hurricane is associated with particular coastal areas, and is characterized by high winds, rainfall, flood impacts, etc.). Taken together, these complementary dimensions mirror the scientist's dual processes of reductionism and synthesis. This structure provides a semantic framework for classifying resources in terms of their underlying knowledge context.

## V. Semantic Web Tools

### *A. Ontology Languages*

An ontology is expressed using a *language* that is typically a specialization of XML. XML is widely supported by existing software tools and is rich enough to express the hierarchical structures inherent in knowledge representation. Resource Description Framework (RDF) is the simplest such ontology language. RDF specializes XML by standardizing meanings for: class, subclass, property, subproperty, domain, range, etc. The DARPA Markup Language (DAML) and DAML+Ontology Inference Layer (DAML+OIL) are further specializations of RDF. These languages add standard meaning for: cardinality, inverse properties, synonyms, and many more concepts. DAML is the most widely used ontology language and is being adopted by W3C as its standard Ontology Web Language (OWL). We adopted DAML for this project, due to its widespread acceptance.

DAML has support for numbers only through a W3C specification [2]. This spec defines number types (e.g., real numbers, unsigned integer) and some abilities to create derivations of these types (e.g. the closed interval between 0 and 1). It contains no operations or relations on these numbers. This is a deficiency, because basic scientific concepts such as "brighter", "higher", "later", or "more northerly" are special cases of the "greater than" relation, when applied in specific domains. This specification also has no notion of a multidimensional space $\mathbf{R}^n$.

The DAML web site [3] maintains libraries of ontologies developed by the community, to enable the work of others to be extended. However, at present there are no ontologies supporting numeric operations (e.g. "greater than" "max"), and only limited support of spatial concepts can be found. A few time ontologies exist, but none took advantage of the fact that time is simply a numerical scale. Therefore, the numerical, space, time, and event ontologies that we develop for SWEET will submitted to the DAML ontology library.

General purpose DAML tools are available from the DAML web site, and include JAVA and Perl parsers, editors, and visualization tools. However, these products did not necessarily support all aspects of the language. None of the editors supported numerical intervals (e.g. the interval [0,1]). This is a deficiency for science applications, where spectral regions are defined in terms of wavelength (e.g. visible light is between 0.3 and 0.7 nanometers), atmospheric layers are defined by altitude (e.g. troposphere is between 0 and 15 km), etc.

### *B. Storage of Ontology Elements*

XML based languages (such as DAML) are well suited to data and model exchange, but are less practical for storage and query of large ontologies. Existing database management systems provide the needed functionality in storage and indexing of robust ontologies, including support for data integrity, concurrency control, etc.

Consequently, we adopted the Postgres object-oriented DBMS to store the names and parent-child relations of our ontology elements. There was no DBMS API available for DAML, so we created two-way translators between the internal DBMS representation and the usual XML representation of the subclass and subproperty relations. By placing all term declarations in the DBMS, any search for terms is very rapid..

### *C. Ontology-Aided Search*

A search tool that is aided by an ontology can potentially locate resources without having an exact keyword match. To verify this claim, we created a search tool that consults the SWEET ontology to find synonymous and more specific terms than those requested. The tool then submits the union of these terms to the GCMD search tool and presents the results. The results verified that additional relevant terms were found from the search, relative to the exact keyword search. During the next stage of this project, we will develop metrics to quantify the increased number of hits.

As a further enhancement to ontology-aided search, we explored methods of automatically discovering associations between terms. For example, the terms "carbon dioxide" and "global warming" would likely be high associated, in the sense that when one term appears, the other is relatively likely. We used the GCMD DIF summaries as the text, from which we created an association matrix. We applied latent semantic analysis (LSA), a method that uses empirical orthogonal functions to find additional hidden associations between terms. We included these association scores in the search tool that we developed. The use of DIF summaries

probably limits the value of this approach (as the summaries are inconsistent in their content). Nevertheless, this exercise showed that additional relevant associated terms could be automatically extracted. In the next stage of this project, we will examine alternatives to LSA that reduce the large computational requirements of processing large matrices.

*D. Application to the ESIP Federation*

Our primary application was data discovery, although numerous other applications are likely to benefit from a greater semantic understanding by information retrieval tools. The ESIP Federation Products and Services Standing Committee plans to help build ontology-based metadata based on each of the EarthRealms. In the process, we expect to give special attention to the different needs of the ESIP Categories across the spectrum of data gathering and distribution, science interpretation and application building, and delivery and usage for public benefit and decision support.

*E. Agents*

We experimented with a network of agents to aid in solving specific types of queries that a user might submit. We assigned an agent to each of four possible query types, that might be representative of user requests.

Each agent is a Perl script that tries to capture what types of problems it can solve directly, what types of problems it can guess a proper response, and what problems it must query the user for further information. The four query types are:

1. Researcher, data specific ("Give me ASTER data for Ecuador, March 1997")
2. Researcher, non-specific ("Give me near-infrared data for the Amazon, 1996")
3. Educated public ("Give me information related to the 1998 El Nino")
4. General ("Show me my house")

Request 1 maps well to the existing ontologies and can be easily satisfied. Request 2 requires prompting user for specific data product. Request 3 requires some assumptions to be made to satisfy request. Request 4 requires significant assumptions to be made to satisfy. In each case, the request is "passed up" the chain until it can be satisfied. In simple cases, we found this approach to be promising. However, much additional agent "learning" must be included for a robust system to be successful on general requests.

*E. Automatic Semantic Tag Creation*

The vision of the semantic web is for data providers to enter tags on their web pages that provide an ontology meaning (or namespace) for technical terms. These namespace tags could be used by search engines to properly classify documents that it indexes. It is unclear whether web page developers will mark up their pages in such a manner, as this requires i) knowledge of the relevant ontologies and ii) time to create the tags. We explored alternate approaches to automatic generation of these tags during the indexing process itself.

Automatic tag creation involves natural language processing to ascertain the meaning of a term based on its context. In some cases, terms have multiple meanings, and tools such as Latent Semantic Analysis (LSA) [4] can be used to distinguish which meaning was intended, based on the appearance of other associated words in the same document.

## VI. Future Research

Much future research is needed to enable the semantic web vision to become a reality. Of particular interest are automation of tasks now being performed manually, including: automatic semantic acquisition, automatic ontology population, and automatic query classification. Accompanying these efforts should be a method of benchmarking, which is necessary to compare our approach with others in the field. There also is a need for better tools for manipulating ontologies. Most of these areas are likely to be addressed by the general ontology community, as they are not specific to the Earth sciences.

The next phase of SWEET is its integration into the Federation Interactive Network for Discovery (FIND). This work will further develop the semantic analysis tools and the agent network and will expand the ontology itself.

## ACKNOWLEDGEMENT

I want to thank Karen Moe and the Earth Science Technology Office for making this effort possible. Also, Michael Pan, Howard Burrows, and Chris Mattmann provided extensive help and support on this project.

## REFERCENES

[1] Global Change Master Directory, "GCMD's Science Keywords and Associated Directory Keywords," http://gcmd.nasa.gov/Resources/valids/index.html

[2] World Wide Web Consortium, "XML Schema Part 2: Datatypes," http://www.w3.org/TR/xmlschema-2.

[3] DAML, "The DARPA Markup Language Homepage," http://www.daml.org/ontologies.

[4] LSA, "Latent Semantic Analysis", http://lsa.colorado.edu.